# Department of Veterans Affairs, University of Utah Consortium Participation in the NLM/AHCPR Large Scale Vocabulary Test

J. Chris Eagon,MD, Eduardo Ortiz,MD, Kenneth A. Zollo,MD, John Hurdle,MD, PhD, Michael J. Lincoln,MD
Departments of Medical Informatics, Internal Medicine, and Surgery
Veterans Affairs IRMFO, VA Medical Center; University of Utah, and Washington University
Salt Lake City, Utah and St. Louis, Missouri

*The Large Scale Vocabulary Test (LSVT) was designed to evaluate how well the Metathesaurus plus planned additions to Meta covered the documentation needs of clinicians. Our consortium collected 10,538 clinical narratives from patient problem lists recorded at 65 Veterans Hospitals, internal medicine ambulatory care practices, diagnostic history and physical examination data elements from Iliad, and nursing shift notes and emergency transport patient records. The results showed 94% of submitted terms resulted in acceptable matches. 49% of submitted terms were judged to be synonymous with the match terms, 35% were judged to be more specific (usually due to modifiers), 2%, were less specific, and 6% had an associative relationship. In 8% of cases either no match was found by the LSVT interface or all proposed matches were rejected by the raters. The LSVT content was quite suitable for coding our narratives. Necessary improvements for an electronic record would include the ability to compose modifiers together with root concepts.*

## INTRODUCTION

The Large Scale Vocabulary Test (LSVT) was intended to determine how well a "combination of existing health related classifications and vocabularies covers vocabulary needed in information systems supporting health care..."[1]. This combination consisted of the current Metathesaurus (Meta) and several planned additions (PA) to Meta, with a matching engine provided by a Web interface. The Department of Veterans Affairs (VA) and University of Utah Health Sciences Center (UUHSC) formed a consortium for this work. This group has previously subcontracted for the Unified Medical Language System (UMLS), developed the Iliad expert system, and produced Meta-based applications for VA's Decentralized Hospital Computer Program (DHCP)[2].

Many coding systems designed for special purposes such as billing are inadequate to code general medical records[3]. Some groups are measuring how well Meta-based systems can code health care records[3,4]. We wished to determine how well the Meta and PA would match our documentation needs. In cases where that match was imperfect or nonexistent, we wished to determine whether future improvements would best be achieved by altering lexical search engines, adding specific terms missing from the dataset, or changing the data model and allowing for a compositional "vocabulary server" approach.

## METHODS

The LSVT involved submitting terms used in local instantiations of the medical record and human evaluation of potential match terms identified by a web-based server. Matches were either exact (requiring the choice or rejection of a single match term), approximate (requiring the choice of 1 out of 20 match terms and a relationship), or not found.

### Subjects

Seventeen clinician raters were recruited, trained, and evaluated for rating consistency[5]. They consisted of seven physicians, two senior medical students, seven nurses, and one senior Ph.D. knowledge engineer. These raters were assigned to rate term sets in their domain of expertise.

### Term Sources

Terms were selected from four sources. First, "Unresolved narratives" (UN) were generated for submission from the VA Problem List application . UNs are clinicians' free text concepts which have failed to match the system's controlled vocabulary[6]. Second, similar UNs were generated by the UUHSC Advanced Clinical Information System (ACIS). Third, nursing terms generated during shift reports at UUHSC[7] and during air transports from multiple centers were selected. Fourth, all history and physical exam concepts from the Iliad 4.5 knowledge base were selected.

**VA Unresolved Narratives (VAUN) (3,500 terms).** All 136,267 UNs received from 65 VA Medical Centers between April 1, 1994 and September 12, 1996 were collected. The 3,614 unique UNs that were generated in three or more copies by two or more sites were then selected for study. 3,500 unique UNs remained after excluding those with only nu-

1091—8280/97/$5.00 

meric data, "v" and "e" ICD codes, or single characters.

**UUHSC ACIS Patient Problems (2,855 terms).** In August, 1996, the ACIS Patient Problem table contained 6,394 UNs. Duplicates, misspellings, multiple concepts, numbers or symbols, and forms of "status post" or "history of" were manually excluded leaving 2,855 narratives for submission.

**Clinical Nursing Data (1,603 terms).** Nursing terms were manually extracted from paper records and corrected for misspellings and abbreviations. Shift notes from medical and surgical floors generated 486 unique terms. The air transport records contained 1,117 terms which occurred in three or more copies.

**Iliad System (3,429 terms).** All 3,429 elements describing history and physical findings were selected for submission to the LSVT.

**Term Aggregation and Submission**

All terms were placed in a master table. The 11,387 terms contained 849 duplicates leaving 10,538 unique terms for submission. These were segmented into files of 50 terms for batch mode submission.

**Rating Procedures**

Each rater underwent one hour of interactive training and a series of two test sets to assure consistent inter-rater agreement on match term and relationship selection. Rules were developed by team consensus to assist raters in making consistent decisions[5]. Exact matches were accepted only if the terms were synonymous. For approximate matches, the raters picked the most closely related match term according to the following preference sequence: synonyms, parents, children, grandparents, great-grandparents, siblings, and other associated terms. Raters selected one of four relationships for every accepted approximate match: synonym, more specific than, broader than, or associated with. When multiple concept narratives were submitted, raters followed special procedures to match constituent concepts.

Because the match rules were liberal for non-synonymous terms, the raters were instructed to comment in every case where the submitted and match terms were not synonymous[5]. A set of 25 coded comments, developed to assist in analysis, fell into five descriptor categories: relationships, modifiers, multiple concepts, exact match disagreements, and miscellaneous comments.

**Analysis**

Upon completion of the rating work, all result files were downloaded into a Microsoft Access table and stratified as follows: All terms (10,538), ACIS (2855), VAUN (3500), NURS (1603), and ILIAD (3429). For each strata, the following parameters were measured: 1) initial LSVT response: the fraction of exact, approximate, and no matches; 2) accepted match rate for exact and approximate matches, and 3) the fraction of submitted terms with the relationships synonym, more specific than, broader than, and associated with.

Terms which failed to match or were rejected by the raters were evaluated by a committee of four raters. A random sample of 200 such terms was selected and problems that caused match failure were repaired as necessary. The repaired terms were resubmitted using both batch and interactive submissions. The analysis included the match relationship, repair type, and term source. Repairs included synonym substitution, spelling correction, abbreviation expansion, modifier truncation, specification of context, and separation of multiple concepts. Some terms were not valid medical concepts and were not resubmitted.

For comment analysis, accepted non-synonymous approximate matches were identified in each strata. Comments were analyzed in each substrata for each of the five comment classes.

## RESULTS

Raters completed the rating of term sets between October 30, 1996 and January 8, 1997. In Table 1, the LSVT interface matching results and the rater acceptance of the matches is listed. The LSVT interface identified an exact match in 15-59% of terms, whereas approximate matches were identified in 37-85% of terms. Rarely, no potential match was identified (0.3-4% of terms). Iliad terms had the lowest exact match rates and nursing terms had the highest rates.

The rate at which matches were accepted by raters was high for all data sets (91-95%). Acceptable match terms were identified in 94% of the 10,330 matched terms. Exact matches were accepted between 94 and 100% of the time with nursing terms having the lowest rate. An approximate match term was accepted between 86 and 95% of the time, with the nursing terms having the lowest percentage.

Table 1: LSVT Interface Response and Match Acceptance

| Term sources | Terms submitted | LSVT Interface Response (% of Submitted terms) | | | Match Acceptability (% of Respective Category) | | | Relationship of Submitted Term to Matched Term (% of Submitted terms) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Exact Match | Approximate Match | Not Found | Overall Match Acceptance | Exact Matches Accepted | Approximate Matches Accepted | Total Synonymous Matches | More Specific Than | Broader Than | Associated With |
| ACIS | 2,855 | 1427 (50) | 1370 (48) | 66 (2) | 2655 (95) | 1407 (99) | 1248 (91) | 1740 (61) | 684 (24) | 36 (1) | 195 (7) |
| VAUN | 3,500 | 1359 (39) | 2062 (59) | 79 (2) | 3215 (94) | 1317 (97) | 1898 (92) | 2212 (63) | 845(24) | 78 (2) | 80 (2) |
| NURS | 1,603 | 949 (59) | 594 (37) | 60 (4) | 1405 (91) | 894 (94) | 511 (86) | 1038 (65) | 171 (11) | 80 (5) | 116 (7) |
| ILIAD | 3,429 | 507 (15) | 2913 (85) | 9 (0.3) | 3262 (95) | 505 (100) | 2757 (95) | 996 (29) | 2001 (58) | 45 (1) | 220 (6) |
| | | | | | | | | | | | |
| Overall | 10538* | 3521 (33) | 6809 (65) | 208 (2) | 9706 (94) | 3413 (97) | 6293 (92) | 5209 (49) | 3661 (35) | 229 (2) | 607 (6) |

*Due to duplicate terms across term sets, the overall total is less than the sum of the individual term sources

95% confidence intervals for all percentages are between +/- 0.13% and +/- 2.1%

The relationships between the submitted terms and the matched terms are indicated in the final four columns of Table 1. In order for exact matches to be accepted, they had to be synonymous with the match term. Adding these terms to the approximate match terms where the synonym relationship was selected, synonymous terms were matched for 49% of the submitted terms. The percentage of synonymous matches was lower for Iliad (29%) than for the other term sources (61-65%). The next most common relationship between submitted term and match term was more specific than, which was selected for 35% of submitted terms (11%-58%).

Table 2 shows the results of manual inspection and resubmission of 200 of the 832 terms where either a match was not found by the LSVT interface, or all possible matches were rejected by the raters. A match was eventually identified in 86% of these terms. The most common term alterations required to identify a match in the case of not found terms were spelling corrections and synonym substitutions. Rejected matches most often required synonym substitutions,

Table 2: Match Failure Repairs

| Random sample of 200 out of 832 terms w here LSVT initially failed to identify appropriate match | | Initial Match Failure Category: Count (% of total in category) | | |
|---|---|---|---|---|
| Final Match Result | Term Alteration Resulting in Closest Match | Not Found | Rejected Match | Total |
| Match Eventually Identified | Synonym Substitution | 12 (22) | 47 (32) | 59 (30) |
| | Spelling Correction | 35 (64) | 16 (11) | 51 (26) |
| | Abbreviation Expansion | 3 (5) | 40 (28) | 43 (22) |
| | Other | 4 (7) | 25 (17) | 29 (15) |
| | No Repair-Rater Error? | 0 (0) | 4 (3) | 4 (2) |
| | Subtotal* | 54 (98) | 119 (82) | 173 (86) |
| No Match Ever Identified | Valid-No Match Found | 0 (0) | 7 (5) | 7 (4) |
| | No Repair-Not Valid | 1 (2) | 19 (13) | 20 (10) |
| | | | | |
| | Total* | 55 | 145 | 200 |

*Since more than one alteration was sometimes required, counts may exceed the total value

For terms where a match was eventually identified, the most common relationships were synonyms (70%) and more specific than (24%). In 14% of match failures, no match was ever identified either because the submitted term was not considered a valid medical concept (10%) or no matching concept could be found in the LSVT dataset despite repeated attempts to identify one (4%).

Using these data, we estimated a range for the overall LSVT interface performance. Assuming all accepted matches are the closest match terms available in the LSVT dataset the maximum performance was: #accepted matches/(total -invalid-never found) = 9,706 / (10,538 – (0.10 x 832) – (0.04 x 832))=93%. The minimum performance assumes that all the non-synonymous accepted matches have closer matches in the LSVT dataset that were not identified by the interface: #synonymous matches/(total-invalid-never found) = 5,209 / (10,538–(0.10 x 832)–(0.04 x 832)) = 50%.

The total concept coverage estimated from these data is high. Using the estimated number of valid submitted terms as the denominator (10,538 – (0.10 x 832) = 10,455), synonymous matches were present in the LSVT dataset for at least 5,209 + (0.70 x 0.86 x 832) = 5,710 or 55%. Matching terms acceptable by our rating rules were present in the LSVT dataset for 9,706 + (0.86 x 832) = 10,421 or 99.7%.

Table 3 shows the frequency of relationship and other comment descriptors for the accepted non-synonymous matches. In all term sets, more specific than was the most common relationship (47-88%) constituting 81% of the overall non-synonymous matches. Associated with matches were next most common (13% overall) followed by broader than (5% overall). Within the terms with the more specific than relationship, 66% were considered a single generation

Table 3: Relationships for Non-synonymous Matches

| Accepted Nonsynonymous Matches | | | Relationship Descriptors: Count (% of n) | | | | | | Modifier Descriptor | Miscellaneous Descriptor | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Term Sources | Relationship | n (% of N) | Child | Parent | Grandchild | Sibling | Multiple Concepts | Undocumented or Other | Total | Abbrev-iation | Mis-spelling |
| ACIS | More Specific Than | 684 (75) | 394 (58) | - | 84 (12) | - | 40 (6) | 166 (24) | 617 (90) | 101 (15) | 28 (4) |
| N = 915 | Broader Than | 36 (4) | - | 28 (74) | - | - | 2 (6) | 6 (17) | 13 (36) | 4 (11) | 0 |
| | Associated With | 195 (21) | 2 | 0 | 25 (13) | 12 (6) | 139 (71) | 29 (15) | 80 (41) | 56 (29) | 17 (9) |
| VAUN | More Specific Than | 845 (84) | 547 (65) | - | 88 (10) | - | 8 (1) | 202 (24) | 782 (93) | 154 (18) | 16 (2) |
| N = 1003 | Broader Than | 78 (8) | - | 65 (83) | - | - | 0 | 13 (17) | 38 (49) | 11 (14) | 4 (5) |
| | Associated With | 80 (8) | 1 | 2 | 1 | 13 (16) | 36 (45) | 27 (34) | 22 (28) | 22 (28) | 4 (5) |
| NURS | More Specific Than | 171 (47) | 71 (42) | - | 14 (8) | - | 35 (20) | 51 (30) | 78 (46) | 0 | 1 (0) |
| N = 367 | Broader Than | 80 (22) | - | 46 (58) | - | - | 3 (4) | 31 (39) | 5 (6) | 0 | 0 |
| | Associated With | 116 (32) | 1 | 0 | 1 | 38 (33) | 41 (35) | 35 (30) | 6 (5) | 1 (1) | 0 |
| ILIAD | More Specific Than | 2001 (88) | 1440 (72) | - | 401 (20) | - | 25 (1) | 135 (7) | 1718 (86) | 2 (0) | 1 (0) |
| N = 2266 | Broader Than | 45 (2) | - | 40 (89) | - | - | 0 | 5 (11) | 14 (31) | 0 | 0 |
| | Associated With | 220 (10) | 7 | 0 | 2 | 53 (24) | 136 (62) | 22 (10) | 62 (28) | 0 | 0 |
| | | | | | | | | | | | |
| Overall | More Specific Than | 3661 (81) | 2428 (66) | - | 582 (16) | - | 106 (3) | 637 (17) | 3160 (86) | 254 (7) | 46 (1) |
| N = 4497 | Broader Than | 229 (5) | - | 172 (75) | - | - | 5 (2) | 52 (23) | 65 (28) | 15 (7) | 4 (2) |
| | Associated With | 607 (13) | 11 | 2 | 28 | 115 (19) | 352 (58) | 99 (16) | 169 (28) | 78 (13) | 20 (3) |

more specific as indicated by the child relationship descriptor, while 16% were two or three generations more specific. In 3%, the submitted term contained multiple concepts. In 17%, coded relationship descriptors were not included in the comment field. The most common reason for this was the fact that some raters did not include relationship descriptors in cases where the only difference between the submitted term and the match term was the presence of a modifier. Most of the undocumented terms would fall under the child relationship classification.

Most of the broader than terms had a parent relationship descriptor as expected. Associated with terms were more heterogeneous than the others, but 19% were described as siblings, and 58% contained multiple concepts. The final columns of Table 3 show that the submitted term and match term often differed by the presence of modifiers, and this was most common in the more specific than matches (86%).

Table 4 shows the frequency of different types of modifiers absent from the match term but present in the submitted term in more specific than matches. Anatomical structure modifiers were most common followed by spatial, temporal, qualitative, and mechanism modifiers.

## DISCUSSION

Our study was limited by at least two factors. First, the selection of term sources was not random so we cannot generalize the results to other segments of the medical record. However, the boundaries of the medical record are not well defined so the degree to which any sample of concepts span the medical record cannot be explicitly stated. We did include sources with a wide range of granularity and degree of modification, and our results confirmed our hypothesis that Iliad terms would have a lower rate of synonymous matches than the other term sets due to the highly modified nature of the terms. A second limitation was that ACIS and nursing data required "cleaning" involving manual review and subjective judgments. This further limits the generalizability of these results. However, VAUN terms had associated information that was used to select frequently used terms from more than one site. Common misspellings

Table 4: Coded Modifier Descriptors for More Specific Than Matches

| More Specific Than Accepted Nonsynonymous Matches | | | Modifier Descriptors: Count (% of n) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Term Sources | n | Coded Modifier Chosen (% of n) | S/P | H/O | R/O | Anatomical structure | Spatial | Temporal | Qualitative | Quantitative | Mechanism | Other |
| ACIS | 684 | 617 (90) | 5 | 2 | 11 | 205 | 129 | 70 | 75 | 28 | 139 | 88 |
| VAUN | 845 | 782 (93) | 89 | 7 | 25 | 219 | 252 | 80 | 70 | 41 | 71 | 77 |
| NURS | 171 | 78 (46) | 0 | 0 | 0 | 11 | 9 | 0 | 13 | 7 | 26 | 16 |
| ILIAD | 2001 | 1718 (86) | 0 | 0 | 4 | 550 | 261 | 450 | 299 | 145 | 167 | 261 |
| | | | | | | | | | | | | |
| Overall | 3661 | 3160 (86) | 94 (3) | 9 (0) | 39 (1) | 974 (27) | 645 (18) | 592 (16) | 453 (12) | 221 (6) | 395 (11) | 441 (12) |

and abbreviations used by clinicians were included in this source, so results here may represent how the LSVT performs against common imperfections in patient problem representation.

Depending on the definition, concept coverage of the LSVT dataset appears to be fairly high. At least 55% of valid submitted terms were present in synonymous form, and using fairly liberal rules, 99.7% of submitted terms were present in a closely related form. Comment analysis of non-synonymous terms indicate that 58-73% of these terms are just one generation away from the submitted term, thus match terms for 75-81% of submitted terms are present in the LSVT dataset within one generation. Previous studies have examined the acceptable match rate in a variety of coding classifications using nursing terms[7], patient problems[4] and a "random" sample of medical record narratives[3]. For patient problems, the closest UMLS 4th ed. match was deemed acceptable in 65%. For the random sample, UMLS v1.3 matches captured the concept completely in 47% and partially in 14%. In our study, the non-synonymous matches may have closer concepts in the LSVT dataset that were not identified by the interface. Determining how often this occurs would allow more accurate assessment of concept coverage and LSVT interface performance, but would require interactive resubmission of terms.

Analysis of non-synonymous matches showed that the inclusion of modifiers is extremely common, and that most modifiers fall into well defined semantic types. Further improvements in the semantic proximity of match terms to submitted terms would be best accomplished by adding the functionality of a compositional grammar. Simply adding more terms will not solve the problem in the long run and will dramatically increase the size of the vocabulary. Others have proposed that some sort of combinatorial strategy would be successful[8], such as an object model or an event definition[9], and our results support this strategy.

The LSVT interface performed well. Synonymous matches present in the LSVT database were presented to the user in 50% of terms, and an acceptable match was identified in 93%. The most common match failures were caused by misspellings, abbreviations, and non-recognition of synonyms, but this occurred in only 8% of submitted terms. Our previous studies with a random sample of Unresolved Narratives indicated that the NLM Knowledge Source Server identified potential matches in only 17% of terms which was similar to the search engine used in DHCP to identify matches in its UMLS-based lexicon[6].

## Acknowledgments

This work was supported in part by the Department of Veterans Affairs, the Washington University Department of Surgery and the National Library of Medicine. We would like to thank Bruce Bray, Helmuth Orthner, Linda Lange, Cheryl Bagley Thompson, and Omar Bouhaddou, who assembled and preprocessed terms from the various sources and acted as raters for this project. We would also like to thank the other raters, Nancy Brazelton, Robert Hausam, Eric Morgan, Cheryl Strong, Kathy Sward, Ben Stevens, Susanne Miller, and Rob Klein, for their diligent work.

## References

1) Humphreys BL, Hole WT, McCray AT, Fitzmaurice JM. Planned NLM/AHCPR large-scale vocabulary test: Using UMLS technology to Determine the extent to which controlled vocabularies cover terminology needed for health care and public health. JAMIA. 1996;3:281-287.
2) Lincoln MJ. Developing and implementing the problem list. In Kolodner RM ed. Computerizing Large Integrated Health Networks: The VA Success. New York, Springer-Verlag 1997; 349-381.
3) Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR, for the Computer-based Patient Record Institute's Work Group on Codes & Structures. The content coverage of clinical classifications. JAMIA. 1996;3:224-231.
4) Campbell JR, Payne TH. A comparison of four schemes for codification of problem lists. SCAMC Proc. 1994:201-5.
5) Ortiz E, Eagon JC, Lincoln MJ. Methods for evaluating inter-rater agreement during the NLM/AHCPR large scale vocabulary test. Submitted Fall AMIA. 1997.
6) Eagon JC, Hurdle JF, Lincoln MJ. Inter-rater reliability and review of the VA unresolved narratives. Fall AMIA Proc. 1996:130-134.
7) Lange LL. Representation of everyday clinical nursing language in UMLS and SNOMED. Fall AMIA Proc. 1996:140-144.
8) Evans DA, Cimino JJ, Hersh WR, Huff SM, Bell DS. Towards a medical concept representation language. The Canon group. JAMIA. 1994;1:207-217.
9) Huff SM, Rocha RA, Bray BE, Warner HR, Haug PJ. An event model of medical information representation. JAMIA. 1995,2:116-134.